Domain Adaptation for Upper Body Pose Tracking in Signed TV Broadcasts

James Charles[1]
j.charles@leeds.ac.uk
Tomas Pfister[2]
tp@robots.ox.ac.uk
Derek Magee[1]
d.r.magee@leeds.ac.uk
David Hogg[1]
d.c.hogg@leeds.ac.uk
Andrew Zisserman[2]
az@robots.ox.ac.uk

[1] School of Computing
University of Leeds
Leeds, UK

[2] Department of Engineering Science
University of Oxford
Oxford, UK

The objective of this work is to estimate upper body pose for signers in TV broadcasts. Given suitable training data, the pose is estimated using a random forest body joint detector (similar to that used by Shotton *et al*. [4], though without requiring depth measurements). However, obtaining such training data can be costly.

The novelty of this paper is a method of transfer learning that is able to harness existing training data and use it for new domains. Our contributions are: (i) a method for adapting existing training data to generate new training data by synthesis for signers with different appearances, and (ii) a method for personalising training data. In addition we add features to the tracker of Pfister *et al*. [3] to improve tracking performance. As a case study we show how the appearance of the arms for different clothing, specifically short and long sleeved clothes, can be modelled to obtain person-specific trackers.

Motivation. The main motivation for tracking the pose is to automatically learn to recognise sign language [1], where upper body layout [2] is of great importance. Tracking is non-trivial due to changing background (the signer is overlaid on the broadcast video, as shown in Figure 1), and also because the signer changes between broadcast and so there are variations in shape and clothing. Furthermore, large quantities of video data are required to learn even a modest number of signs, implying the tracker has to be reliable over long video sequences and for multiple signers, and require little or no human supervision.

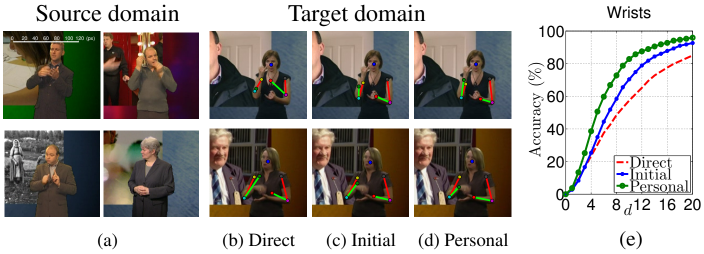
Summary of method. The tracker in our previous work [3] was only designed for application on a domain of signers wearing long sleeved clothing as shown in Figure 1(a). Our target domain is signers wearing short sleeved clothes, shown in Figure 1(b). We generate training data for the target domain from a source domain of long sleeved clothes in two stages. Stage 1 of our method (output shown in Figure 2) uses material from the source domain to generate semi-synthetic training data of signers wearing sleeves of a specific length. With this, one can retrain multiple general upper body pose trackers for a particular sleeve length, as shown in Figure 1(c). Stage 2 (shown in Figures 3 and 4) re-synthesises the training data to become signer specific. Then the refined semi-synthetic data is used to train a personalised tracker tuned to a particular person's arm shape and sleeve length, as shown in Figure 1(d). Each step contributes a significant boost to pose estimation performance, as shown in Figure 1(e).

[1] P. Buehler, M. Everingham, and A. Zisserman. Learning sign language by watching TV (using weakly aligned subtitles). In *Proc. CVPR*, 2009.

[2] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Upper body detection and tracking in extended signing sequences. *IJCV*, 2011.

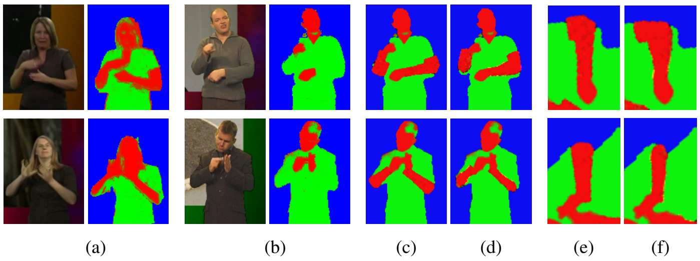
[3] T. Pfister, J. Charles, M. Everingham, and A. Zisserman. Automatic and efficient long term arm and hand tracking for continuous sign language TV broadcasts. In *Proc. BMVC*, 2012.

[4] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proc. CVPR*, 2011.



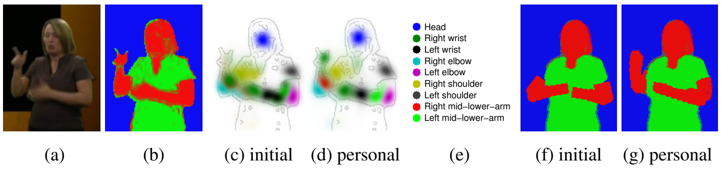
(a) (b) Direct (c) Initial (d) Personal (e)

Figure 1: **Training an upper body joint detector using knowledge from the source domain.** In this example, the transfer is from a source domain (a) where signers wear long sleeves to a target domain (b)–(d) where signers wear short sleeves. The red and green lines on the signer show the output pose. The target domain results show: (b) the detector trained directly from the source domain; (c) trained from initial semi-synthetic images constructed using the source domain; and (d) from personal semi-synthetic images constructed from both source and target domain. (e) Accuracy of (b)–(d) in detecting the wrist joint in the target domain. Accuracy is percentage of predicted joints within a distance *d* pixels from ground truth (scale bar shown in top left of (a)). Note the improvement brought by the two stages – for example the accuracy almost doubles at 8 pixels.



(a) (b) (c) (d) (e) (f)

Figure 2: **Stages in synthesising training data.** Rows depict different sleeve length (sleeves in the top row are shorter than in the bottom row). (a) Raw RGB image and CP image counterpart of *short* sleeve signers. (b) Example *long* sleeve signers and CP images. (c) Synthetically produced CP images of short sleeves using CP images from (b) and of initial arm templates. (d) Personalised synthesis using learnt arm templates. For closer comparison, rotated left arms of the synthetic images in (c) and (d) are shown in (e) and (f) respectively.



(a) (b) (c) initial (d) personal (e) (f) initial (g) personal

Figure 3: **Illustration of sample and verify procedure.** (a) Input signer and (b) CP image. (c) Shows body joint confidence map from initial forest, (d) confidence map produced with a personalised forest and (e) the body joint colour key. Most likely whole-image templates using initial (f) and personalised (g) arm templates, computed by sampling joint locations from (c) and (d) respectively. The personalised model produces both better joint samples (most notably for right wrist) and likelihood function.

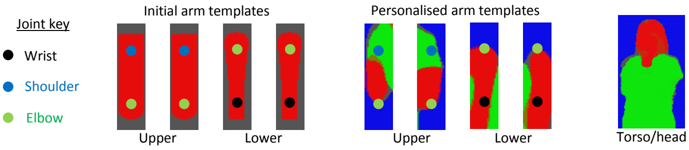


Figure 4: **Initial and personalised arm templates**. The personalised arm and torso/head templates are for the signer in Figure 3(a).